NOTES 10: SAMPLING DISTRIBUTIONS

Stat 120 | Fall 2025

Prof Amanda Luby

"Big picture" picture:

Quantity	Statistic	Parameter
Mean		
Proportion		
Standard Deviation		
Correlation		
Regression Coefficient		

Carleton publishes an "at a glance" page with some facts and figures about the student body: https://www.carleton.edu/about/carleton-at-a-glance/

Some highlights:

- Geographic distribution:
  - Midwest 36.7%
  - West 21.7%
  - East 17.3%
  - South 10.9%
  - International 11.8%
  - Other 1.6%
- 34% BIPOC
- 12% are among the first generation in their families to attend college
- 61% graduated in the top 10% of their high school class

In a moment, we're going to see one of these quantities for our class. Before we do, what is your *best guess* for each of these quantities?

Example: In this set-up, what is the:

- Population
- Sample
- Parameter

- Statistic

We know that our class will likely not have **exactly** 36.7% from the Midwest, but we probably wouldn't expect it to be 0% or 90%.

> Sampling variability

We might start to ask ourselves, what if a *different* set of 32 students enrolled in this course?

First, we create a population.

```
# A tibble: 2,007 x 2
   student_id midwest
        <int> <chr>  
 1         72 Yes    
 2       1327 No     
 3       1100 No     
 4        244 Yes    
 5        308 Yes    
 6        152 Yes    
 7        676 Yes    
 8       1397 No     
 9        518 Yes    
10       1247 No     
# i 1,997 more rows
```

Then, we take a random sample:

```
set.seed(100424)
sample1 <- carls |>
  sample_n(32)
sample1
```

```
# A tibble: 32 x 2
   student_id midwest
        <int> <chr>  
 1       1826 No     
 2       1997 No     
 3        279 Yes    
 4       1505 No     
 5        770 No     
 6       1111 No     
 7       1309 No     
 8       1912 No     
 9       1302 No
10       1959 No
# i 22 more rows
```

and calculate the proportion of “yes” responses:

```
sample1 |>
  group_by(midwest) |>
  summarize(
    n = n()
  ) |>
  mutate(p_hat = n/sum(n))
```

```
# A tibble: 2 x 3
  midwest     n p_hat
  <chr>   <int> <dbl>
1 No         21 0.656
2 Yes        11 0.344
```

This isn’t *super* useful, but if we do it a bunch of times, we can start to see what a range of possible samples could look like. (*Note:* this code requires the {`infer`} package)

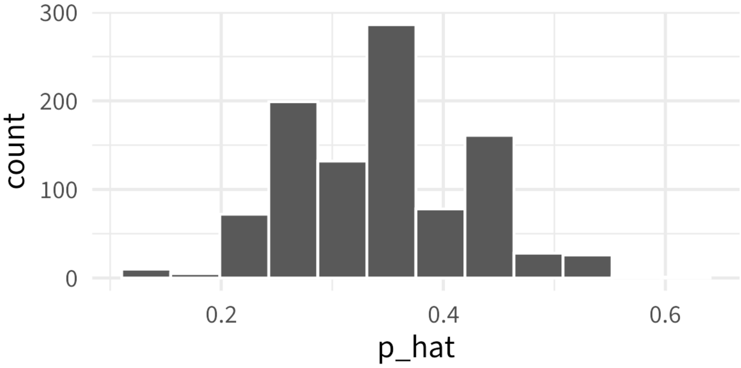
```
many_samples <- carls |>
  rep_sample_n(35, reps = 1000, replace = TRUE) |>
  group_by(replicate, midwest) |>
  summarize(
    n = n()
  ) |>
  mutate(p_hat = n/sum(n)) |>
  filter(midwest == "Yes")

many_samples
```

```
# A tibble: 1,000 x 4
# Groups:   replicate [1,000]
   replicate midwest     n p_hat
       <int> <chr>   <int> <dbl>
 1         1 Yes         9 0.257
 2         2 Yes         7 0.2
 3         3 Yes        11 0.314
 4         4 Yes        15 0.429
 5         5 Yes        18 0.514
 6         6 Yes         7 0.2
 7         7 Yes        18 0.514
 8         8 Yes        14 0.4
 9         9 Yes        17 0.486
```

```
10        10 Yes        15 0.429
# i 990 more rows
```

Looking at this first few rows, we can start to get a sense of the range of possible sample proportions, but there are 990 rows that we can't see. Let's make a graph!



Example: Carleton Mission Statement

In your own words: provide explanations for:

Population distribution

Sample distribution

Sampling distribution